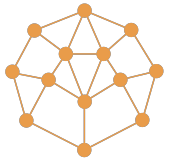
#1 Cleaning & EDA

by Saturdays.AI

Saturdays.AI

Machine Learning

Week 1



Schedule

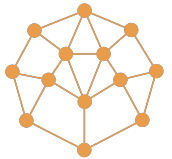
State of the course

Session 1 Review

Challenge

Notebook + resources

State of the course



#1 Cleaning & Exploratory Data Analysis 🔴 Today!

#2 Regression & Support Vector Machine (SVM) ➡️ SOON

#3 Decision Trees & Random Forest Deep Dive ➡️ SOON

#4 Unsupervised Learning + Clustering ➡️ SOON

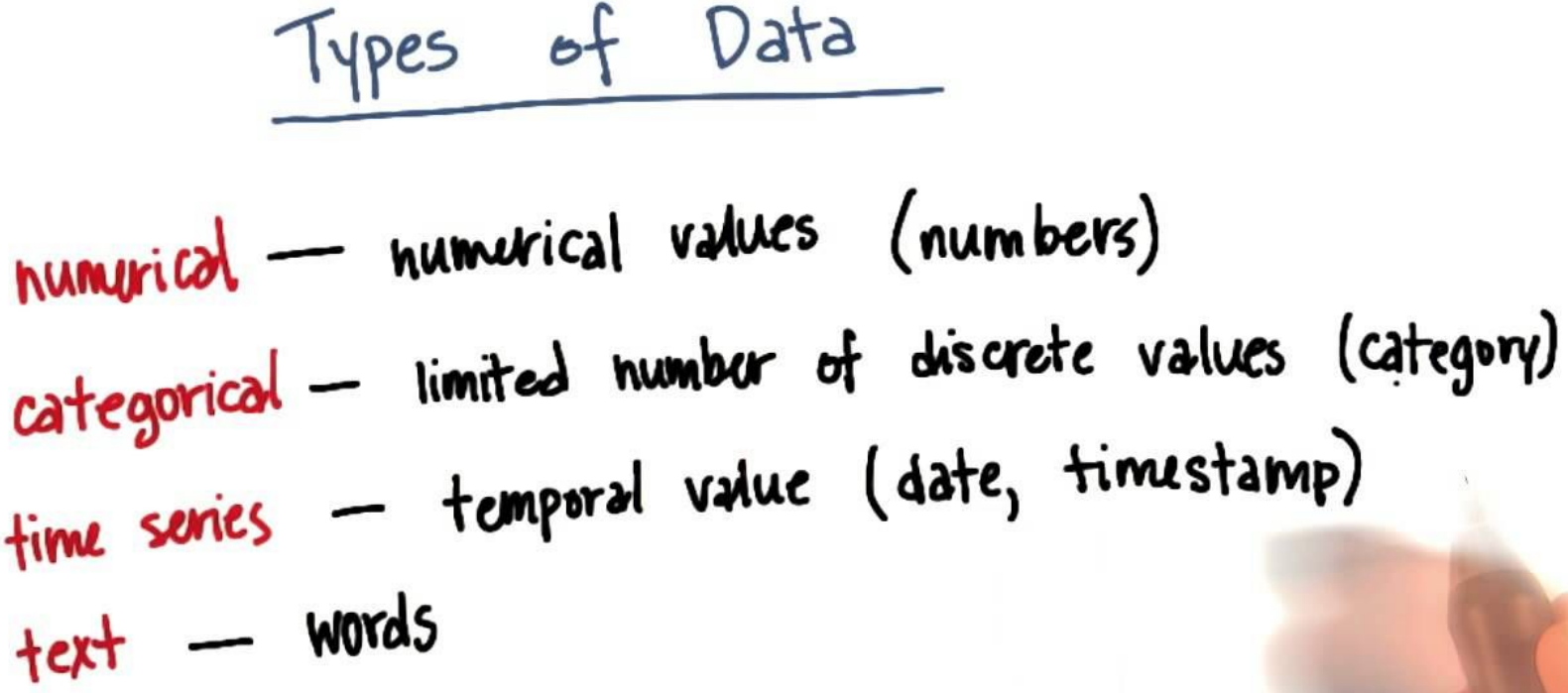
#5 Time Series Analysis + Data Viz ➡️ SOON

#6 Neural Networks, Gradient Descent ➡️ SOON

#7 NLP ➡️ SOON

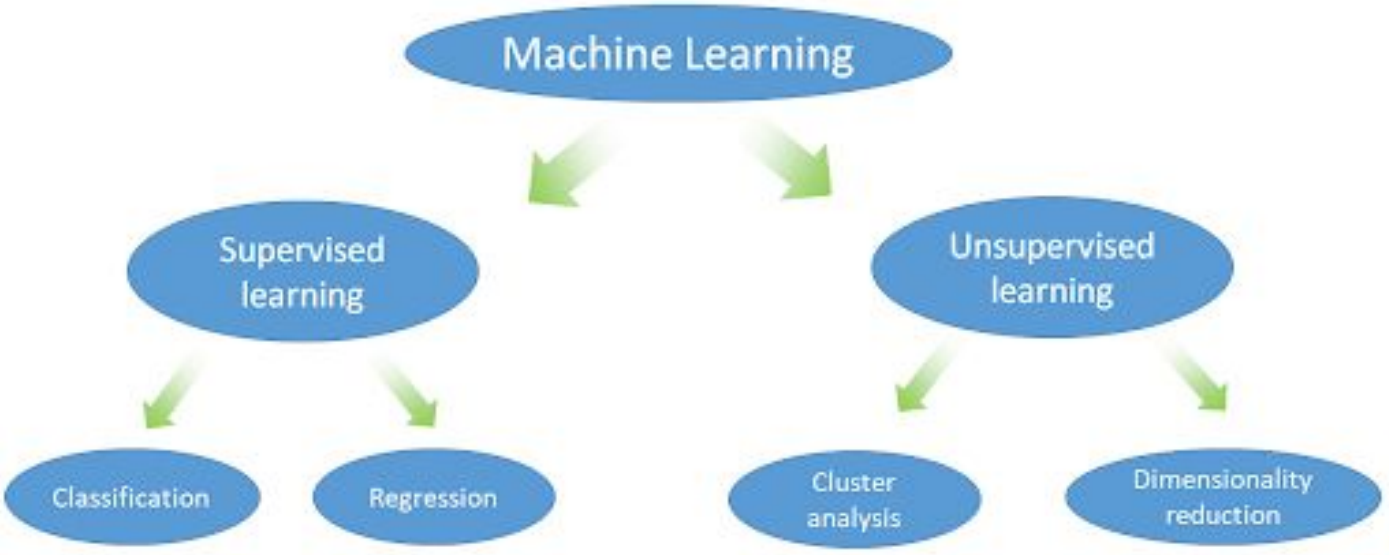
ML: Data Types

Saturdays.AI



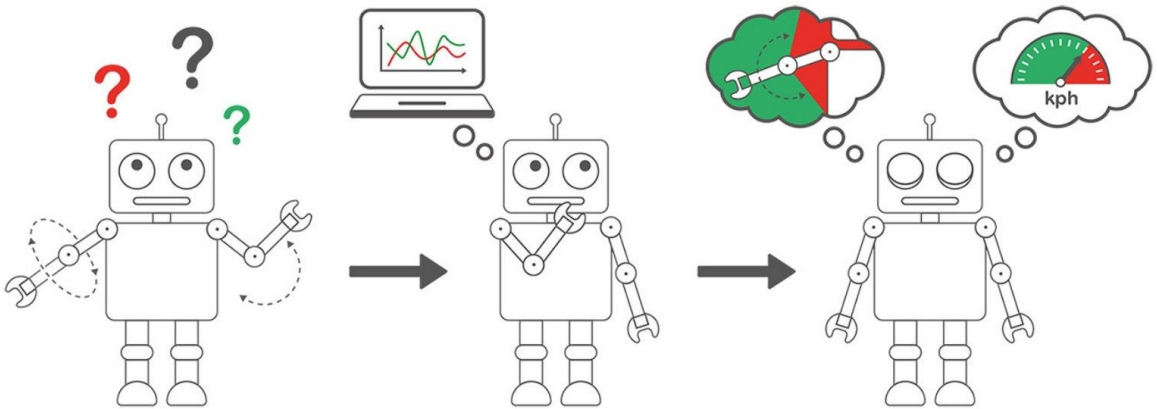
ML: Decision Types

Saturdays.AI



ML: Algorithm architecture types

Saturdays.AI



IS THAT ALL THERE IS ?

ML is interdisciplinary

Machine learning is …

- Computer science + statistics + mathematics
- The use of data to **answer questions**

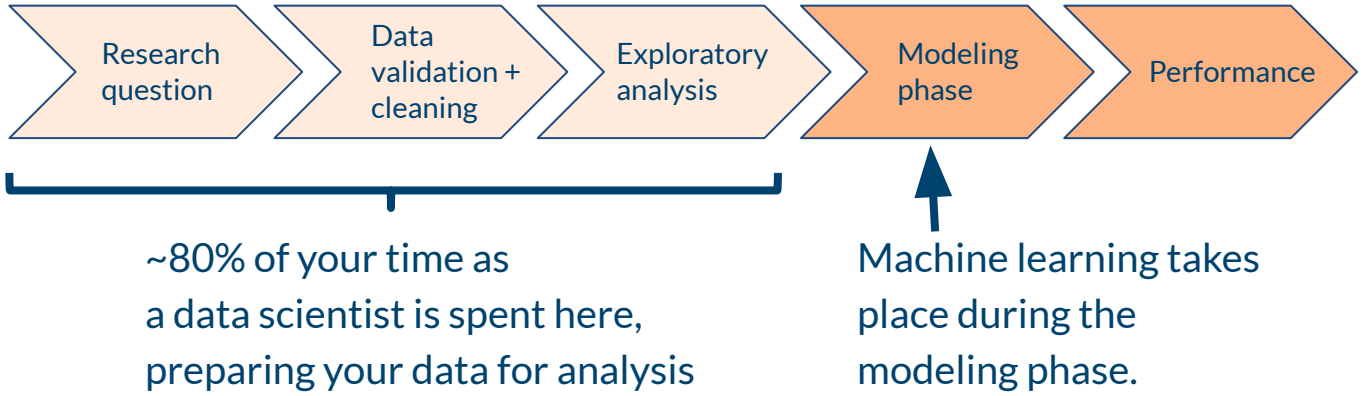
Critical thinking combined with technical toolkit

Source: deltanalytics.org

ML help us answer questions

Saturdays.AI

- **How do we define the question?**
  Before we even get to the models/algorithms, we have to learn about our data and define our research question.



Source: deltanalytics.org

Research question

Examples of research questions:

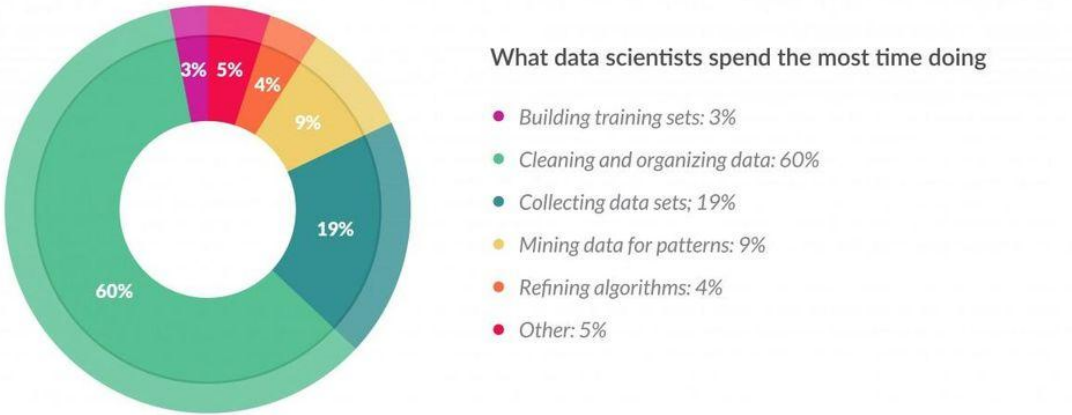
- Does this patient have malaria?
- Can we monitor illegal deforestation by detecting chainsaw noises in audio streamed from rainforests?

Source: deltanalytics.org

Data Validation and Cleaning

Saturdays.AI

"Data preparation accounts for about 80% of the work of data scientists."



Source: Survey of 80 data scientists. Forbes article, March 23, 2016.

Data Cleaning

Saturdays.AI

Why do we need to validate and clean our data?

Data often comes from multiple sources

- Do data align across different sources?

Data is created by humans

- Does the data need to be transformed?
- Is it free from human bias and errors?

Source: deltanalytics.org

Data Cleaning

spreadsheet = dataset

1 column = 1 attribute

target variable

	A	B	C	D	E	F	G	H	I
1	id	Name	Surname	Age	Height	Why interested	Date	Organization	Student
2	995234	Marc	Fossatti	30	1,56	Use in a project	2019-11-04 7:57:38	11	1
3	1249609	Julia	Nicolari	23	1,67	e my knowledge and develop a	2018-05-04 14:13:01	4	1
4	1385554	Pol	Martinez	35	1,52	on how it could be applied on i	2019-11-06 20:42:08	16	1
5	2543328	Martina	Rochon	43	1,54	/ and apply AI to healthcare ir	2019-11-03 20:02:56	95	0
6	3326849	Emma	Silva	20	1,75	ity and get step forward in my	2018-03-20 12:07:09	91	0
7	3588497	Alex	Beloqui	35	1,6	pply AI to solve civil engineeri	2019-11-04 13:44:57	12	1
8	1987304	Jan	Schwarz	29	1,82	my career by learning about r	2019-10-30 19:34:08	1	0
9	1455322	Maria	Sosa	30	1,59	e future and I love that the ma	2019-11-07 13:32:51	11	1
10	1247369	Nil	De maria	22	1,7	e able to do with AI and I wou	2019-11-03 20:49:56	2	0
11	3593956	Leo	Hernandez	41	1,55	re about the future and to get	2018-04-09 14:50:01	87	1
12	3449648	Eric	Nuñez	28	1,57	for knowledge	2019-11-05 14:25:31	68	0
13	1033368	Enric	Bonilla	52	1,78	I am interested in NLP	2019-10-31 7:53:17	1	0
14	1178833	Pau	Seade	27	1,63	se of the opportunities you ca	2019-11-05 17:35:27	11	1
15	655000	Marti	Ferrari	26	1,85	ou can do with data and predi	2019-10-31 10:04:33	1	0
16	3877670	Lucia	Madera	34	1,62	is world a better place howev	2018-03-09 0:44:10	97	0
17	1679960	Paula	Martinez	33	1,73	ated to my field of expertise, I	2019-11-04 20:05:55	94	1
18	3205255	Hugo	Perez	34	1,58	:) BarcelonaTECH where I hav	2019-10-30 19:38:31	94	1
19	1793175	Biel	Pereira	23	1,54	technology that will be a gam	2019-11-06 9:42:45	16	1
20	1250238	Laia	Sosa	24	1,65	l all the improvements that ca	2019-11-03 19:26:21	16	1
21	3694615	Sofia	Cabrera	29	1,5	nalyst and I want to improve i	2018-03-16 22:22:54	88	0
22	1887660	Lucas	Gutierrez	30	1,52	AI in order to make an impac	2019-11-11 17:15:52	4	0
23	1704998	Aïna	Prospero	41	1,73	Because it's awesome	2019-11-03 22:05:54	65	1

1 row = 1 observation

training data

Saturdays.AI

Data Cleaning

Saturdays.AI

Data cleaning involves identifying any issues with our data and confirming our qualitative understanding of the data.

Missing Data
Is there missing data? Is it missing systematically?

Data Type
Are all variables the right type?
Is a date treated like a date?

Times Series Validation
Is the data for the correct time range?
Are there unusual spikes in the volume of loans over time?

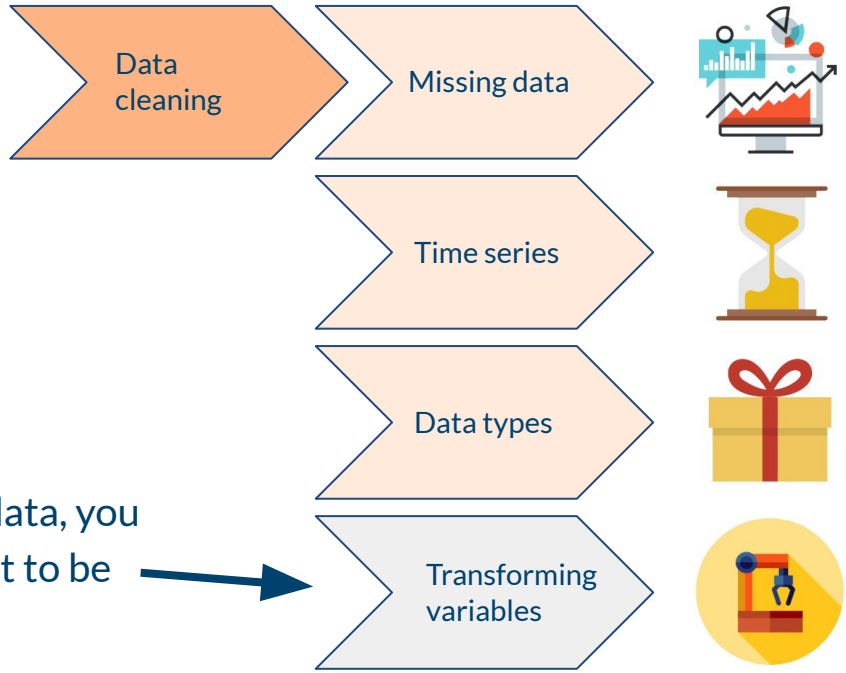
Data Range
Are all values in the expected range?
Are all loan_amounts greater than 0?

Source: deltanalytics.org

Data Cleaning

Saturdays.AI

Let's step through some examples:



After gaining an initial understanding of your data, you may need to transform it to be used in analysis

Source: deltanalytics.org

Data Cleaning: Missing data

Saturdays.AI

Very few datasets have no missing data; most of the time you will have to deal with missing data.

The first question you have to ask is what type of missing data you have.

Missing completely at random:
no pattern in the missing data. This is the best type of missing you can hope for.

Missing at random:
there is a pattern in your missing data **but not** in your variables of interest.

Missing not at random:
there is a pattern in the missing data that systematically affects your primary variables.

Source: deltanalytics.org

Data Cleaning: Missing data

Saturdays.AI

Sometimes, you can replace missing data.

- Drop missing observations
- Populate missing values with average of available data
- Impute data:
  - Educated Guessing
  - Average Imputation
  - Common-Point Imputation
  - Regression Substitution
  - Multiple Imputation

What you should do depends heavily on what makes sense for your research question, and your data.

Lecture: 7 Ways to Handle Missing Data

Source: deltanalytics.org

Data Cleaning: Time series

If we have observations over time, we need to do time series validation.

Ask yourself:

a. Is the data for the correct time range?

b. Are there unusual spikes in the data over time?

What should we do if there are unusual spikes in the data over time?

Source: deltanalytics.org

Data Cleaning: Time series

Saturdays.AI

How do we address unexpected spikes in our data?

Data anomaly

Systematic spike

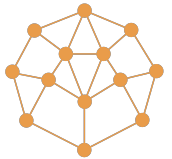
For certain datasets, (like sales data) systematic seasonal spikes are expected. For example, around Christmas we would see a spike in sales venue. This is normal, and should not necessarily be removed.

Random spike

If the spike is isolated it is probably unexpected, we may want to remove the corrupted data.
For example, if for one month sales are recorded in £ rather than €, it would corrupt the sales figures.
We should do some data cleaning by converting to € or perhaps remove this month.

Source: deltanalytics.org

Data Cleaning: Data types



Are all variables the right type?

Many functions in Python are type specific, which means we need to make sure all of our fields are being treated as the correct type:

	integer	float	string	date
	loan_amount	**partner_id**	**sector**	**posted_date**
1957	50	156.0	Personal Use	2017-04-11
78437	350	133.0	Clothing	2013-08-07
116723	575	156.0	Agriculture	2011-01-04

Lecture: Datacarpentry - Data types & formats

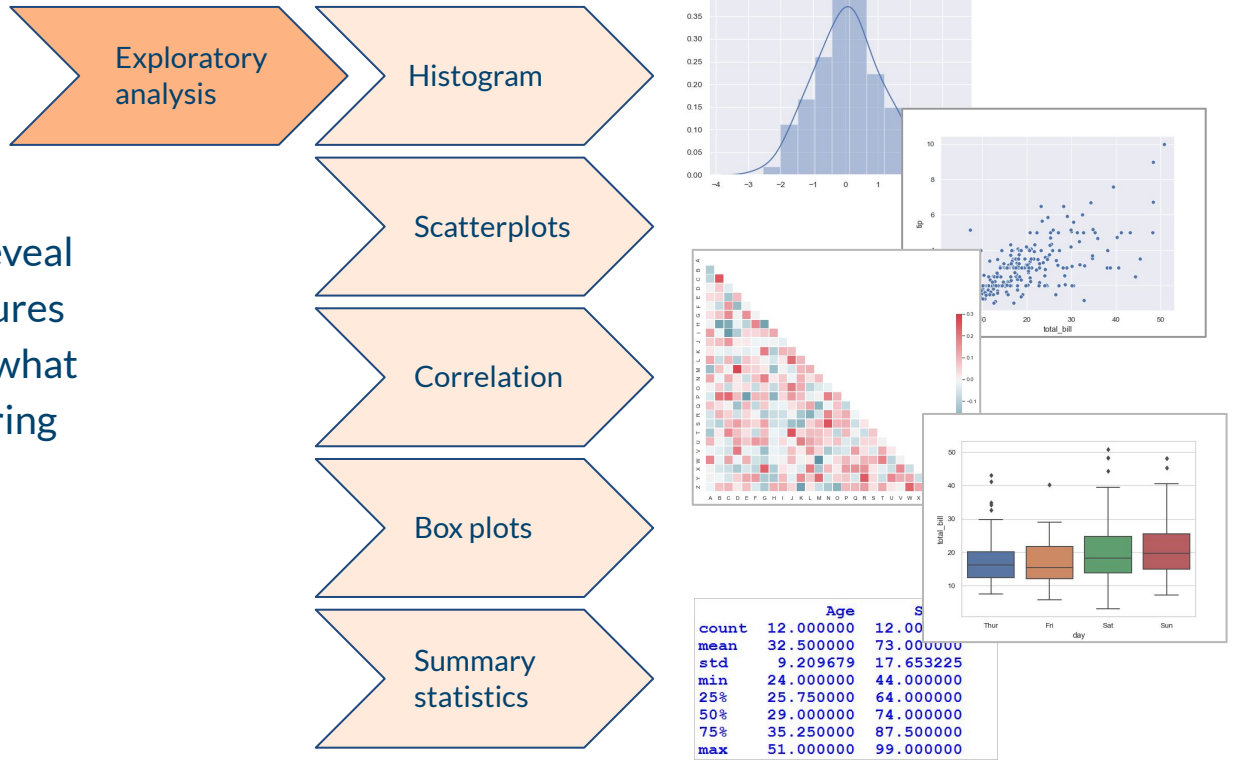
Source: deltanalytics.org

Exploratory analysis

Saturdays.AI

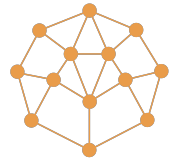
The goal of exploratory analysis is to better understand your data.

Exploratory analysis can reveal data limitations, what features are important, and inform what methods you use in answering your research question.



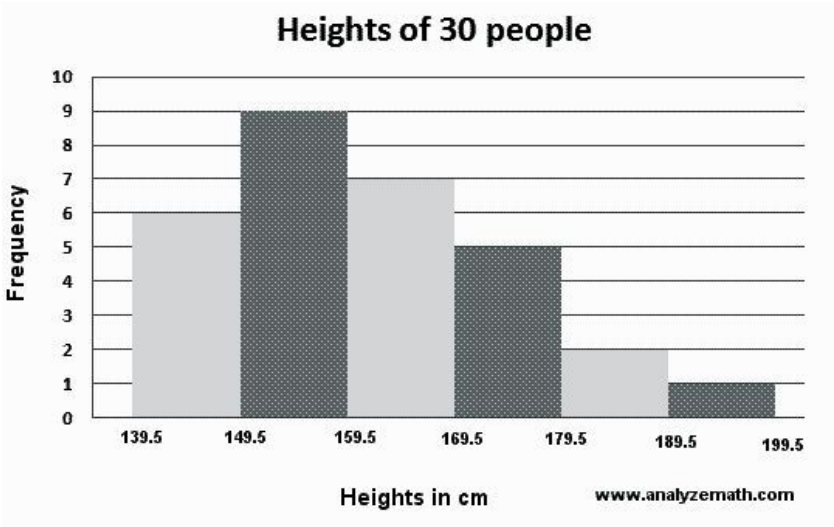
Source: deltanalytics.org

Exploratory analysis: Histogram



Histograms tell us about the distribution of the feature.

A histogram shows the frequency distribution of a continuous feature.
Here, we have height data of a group of people. We see that most of the people in the group are between 149 and 159 cm tall.



Lecture: University of Florida, Histograms & Stemplots

Source: deltanalytics.org

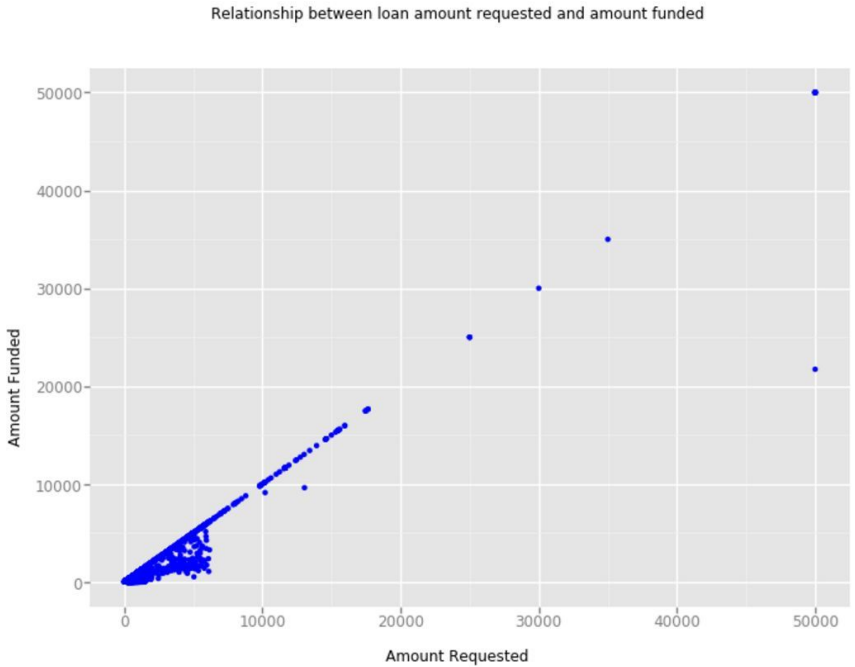
Exploratory analysis: Scatterplot

Saturdays.AI

Scatter plots provide insight about the relationship between two features.

Scatter plots visualize relationships between any two features as points on a graph.

They are a useful first step to exploring a research question. Here, we can already see a positive relationship between amount funded and amount requested.
What can we conclude?



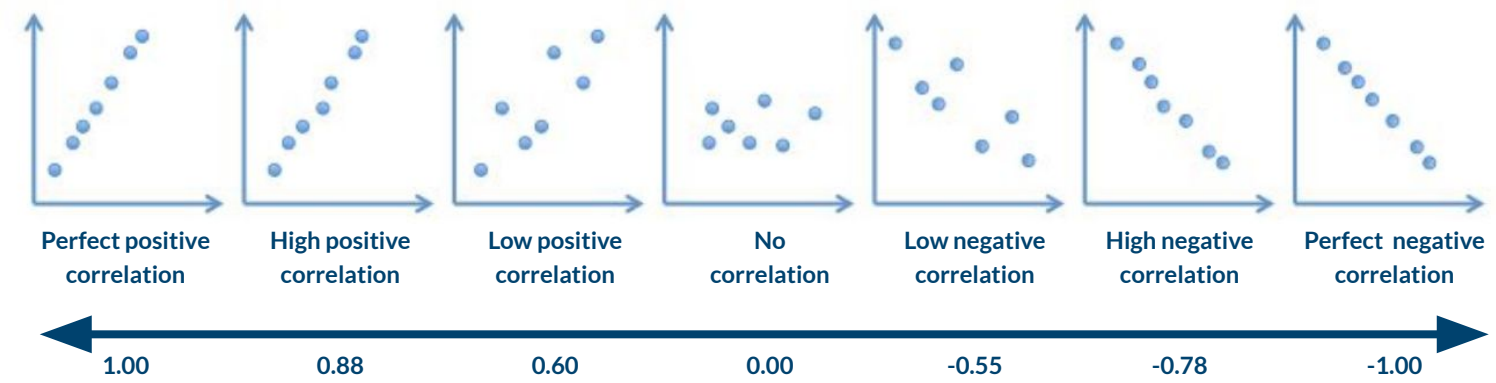
Lecture: University of Florida, Scatterplots

Source: deltanalytics.org

Exploratory analysis: Correlation

Saturdays.AI

Correlation is a useful measure of the strength of a relationship between two variables. It ranges from -1.00 to 1.00



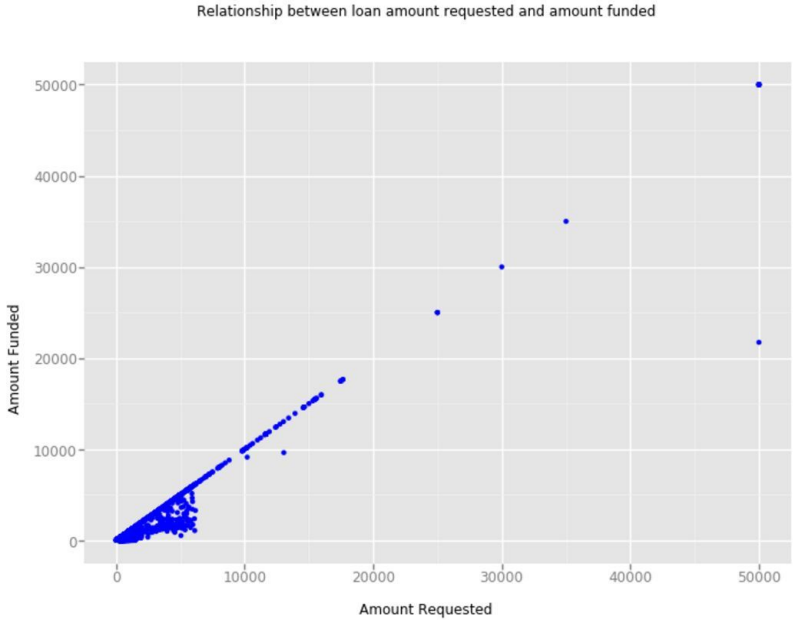
Go further with this fun game.

Source: deltanalytics.org

Exploratory analysis: Correlation

Saturdays.AI

Correlation does not equal causation



Correlation: 0.96

If you wanted to be funded, and were presented with this graph only, you might conclude that it is a good idea to request 50k €.

But common sense tells us that this conclusion doesn't make a lot of sense. Just because you request a lot doesn't mean you will be funded a lot!

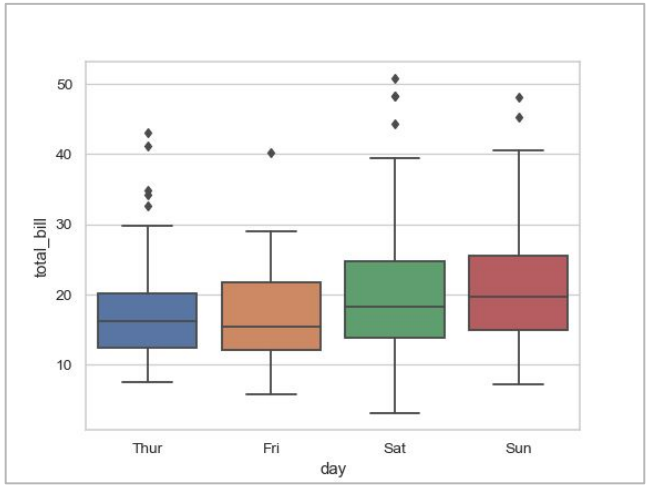
Lecture: University of Florida, Causation

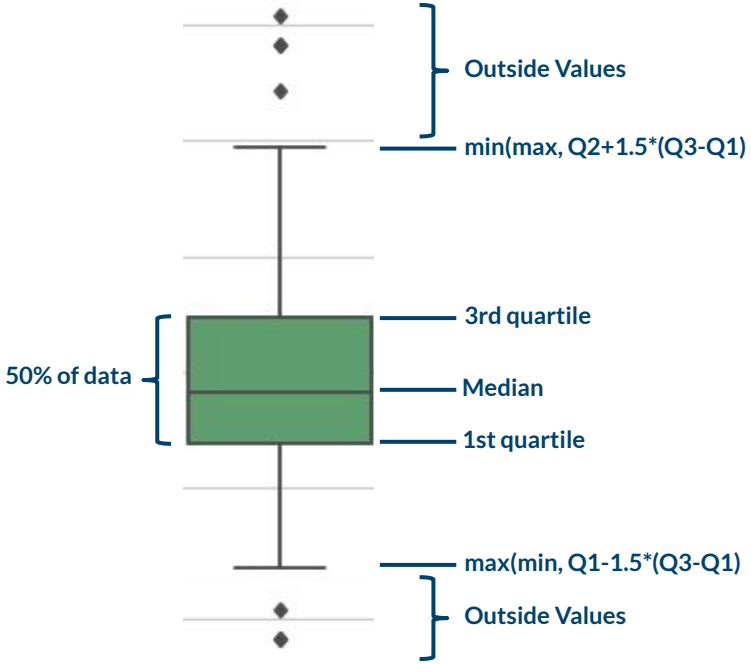
Source: deltanalytics.org

Exploratory analysis: Boxplot

Saturdays.AI

Boxplots are a useful visualization of certain summary statistics.



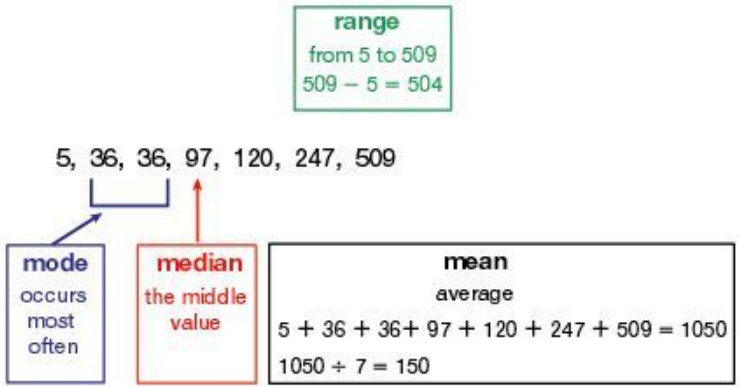


Lecture: University of Florida, Boxplot

Exploratory analysis: Summary statistics

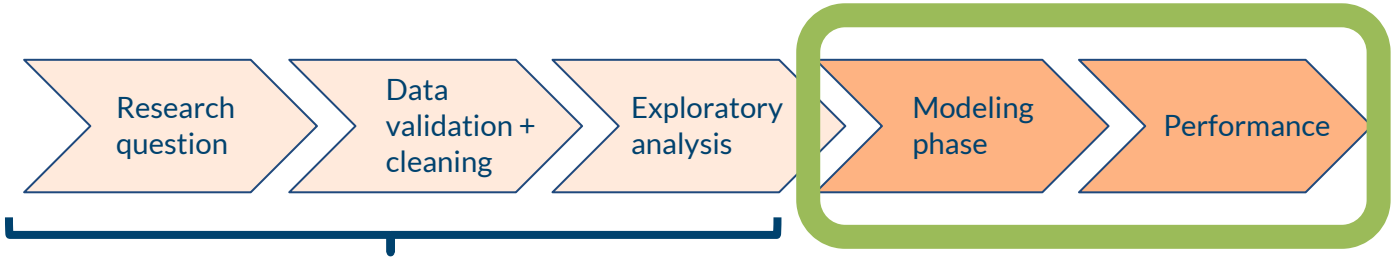
Saturdays.AI

Mean, median, frequency are useful summary statistics that let you know what is in your data.



Source: deltanalytics.org

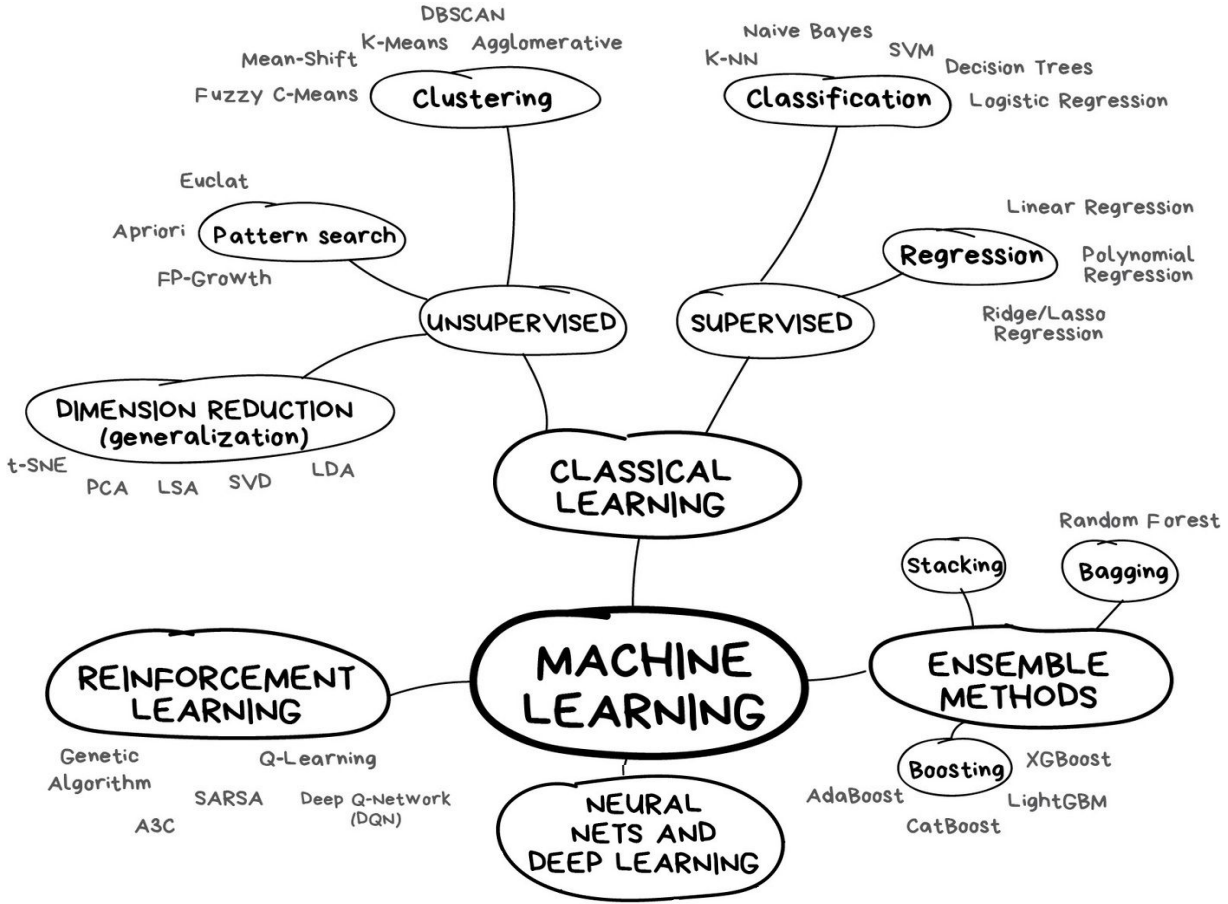
Machine learning helps us answer questions



NOW!

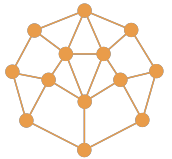
Source: deltanalytics.org

ML Algorithms



Saturdays.AI

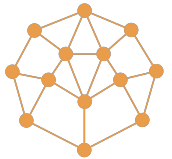
Supervised learning



Basically SUPERVISED LEARNING is used when our data is labelled. This means that we have an output variable for all the input variables.
For example, we have the following dataset:

# chol	# fbs	# restecg	# thalach	# exang	# oldpeak	# slope	# ca	# thal	# target
223	0	1	169	0	0	2	4	2	1
220	0	1	144	0	0.4	1	4	3	1
175	0	1	173	0	0	2	4	2	1
175	0	1	173	0	0	2	4	2	1
247	1	0	143	1	0.1	1	4	3	0
231	0	1	146	0	1.8	1	3	3	1
233	1	1	147	0	0.1	2	3	3	1
246	1	0	173	0	0	2	3	2	1
286	0	0	108	1	1.5	1	3	2	0
225	0	0	114	0	1	1	3	3	0
216	0	0	131	1	2.2	1	3	3	0

Supervised learning



The majority of practical machine learning uses supervised learning.

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

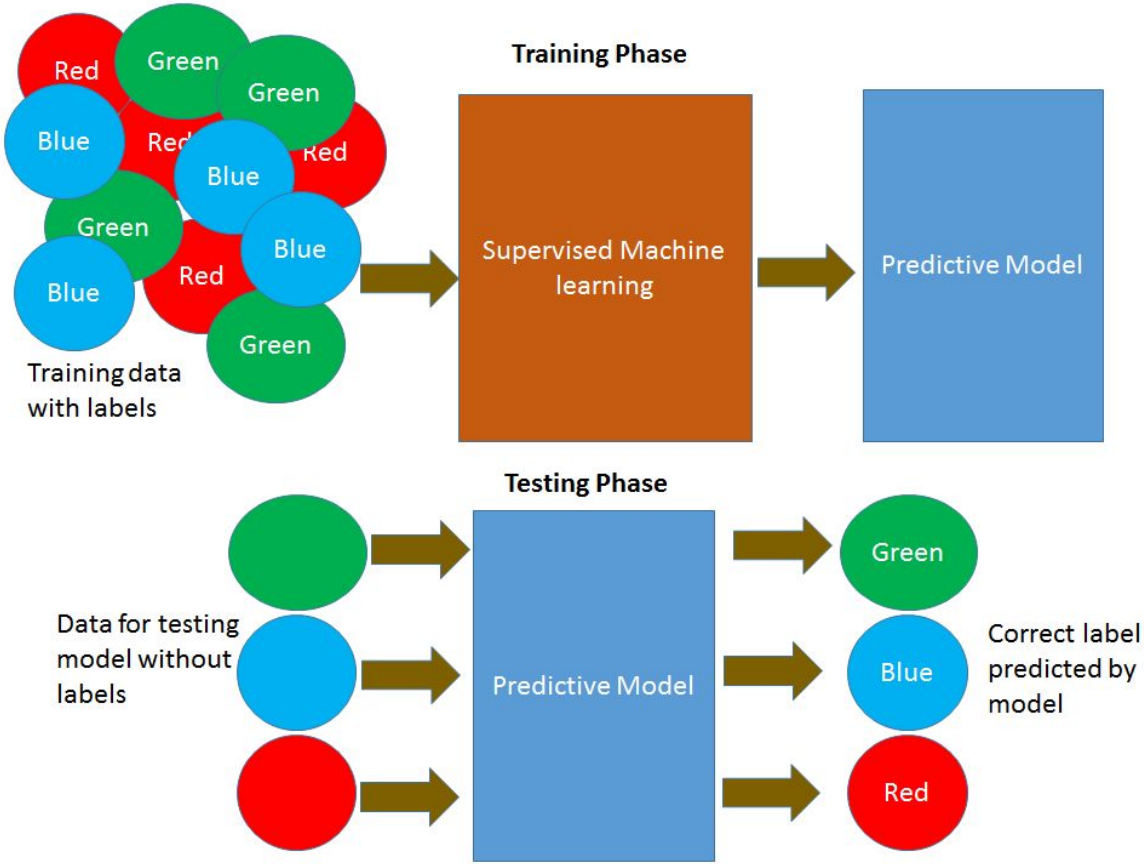
$$Y = f(X)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

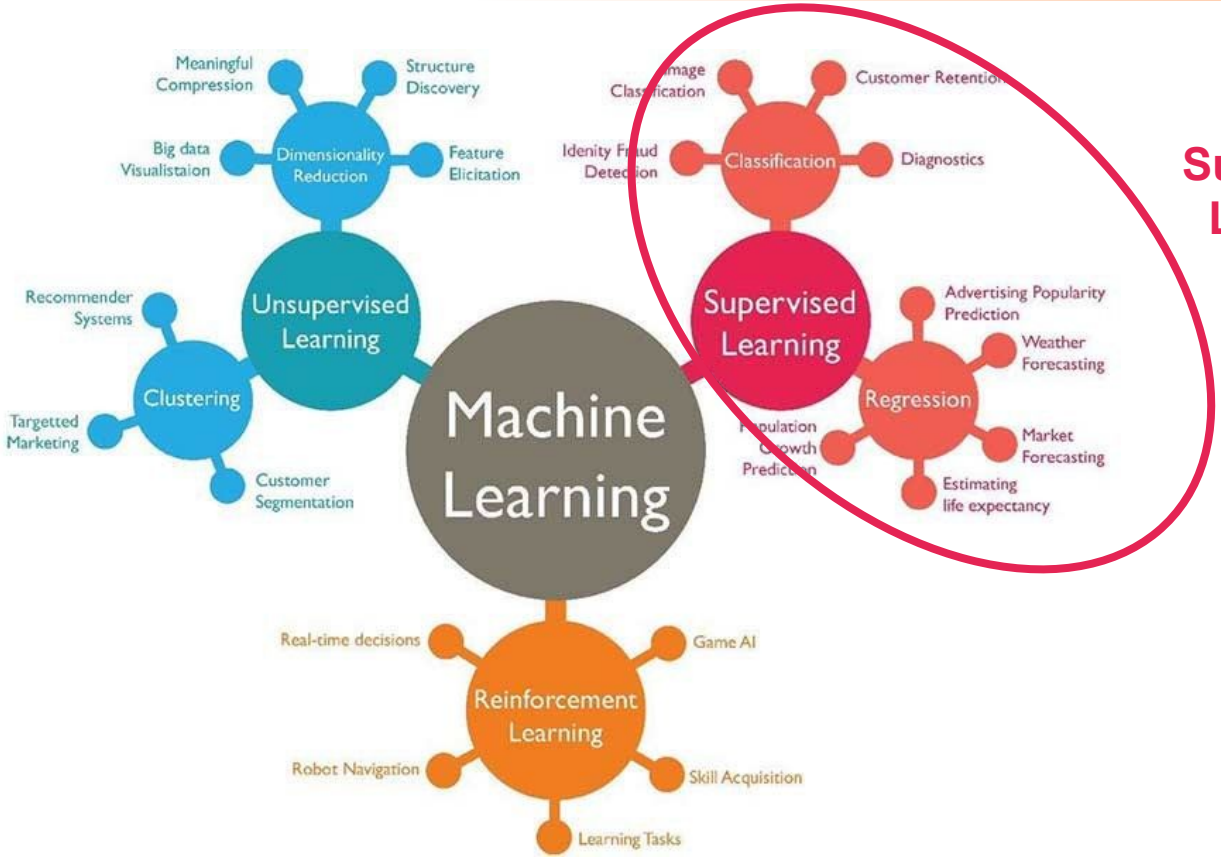
It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Supervised learning

Saturdays.AI



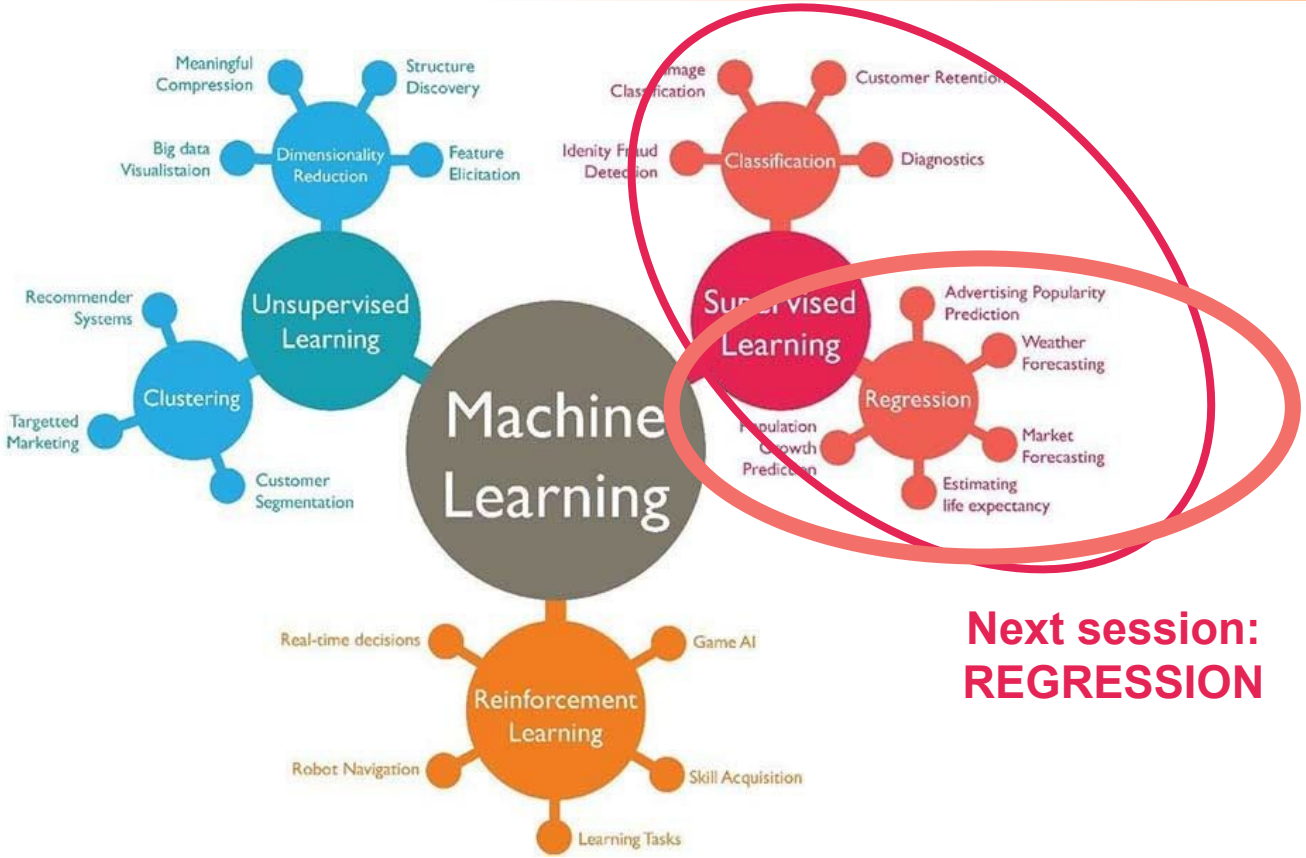
Machine learning



Supervised Learning

Saturdays.AI

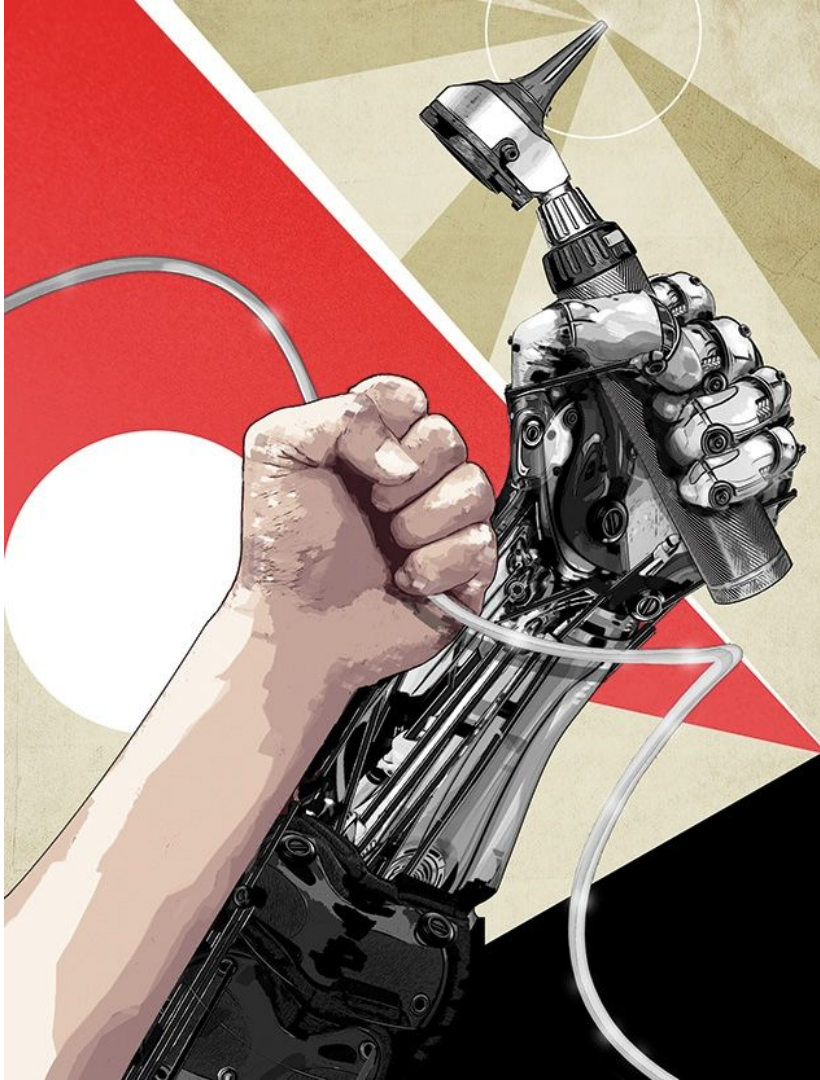
Machine learning



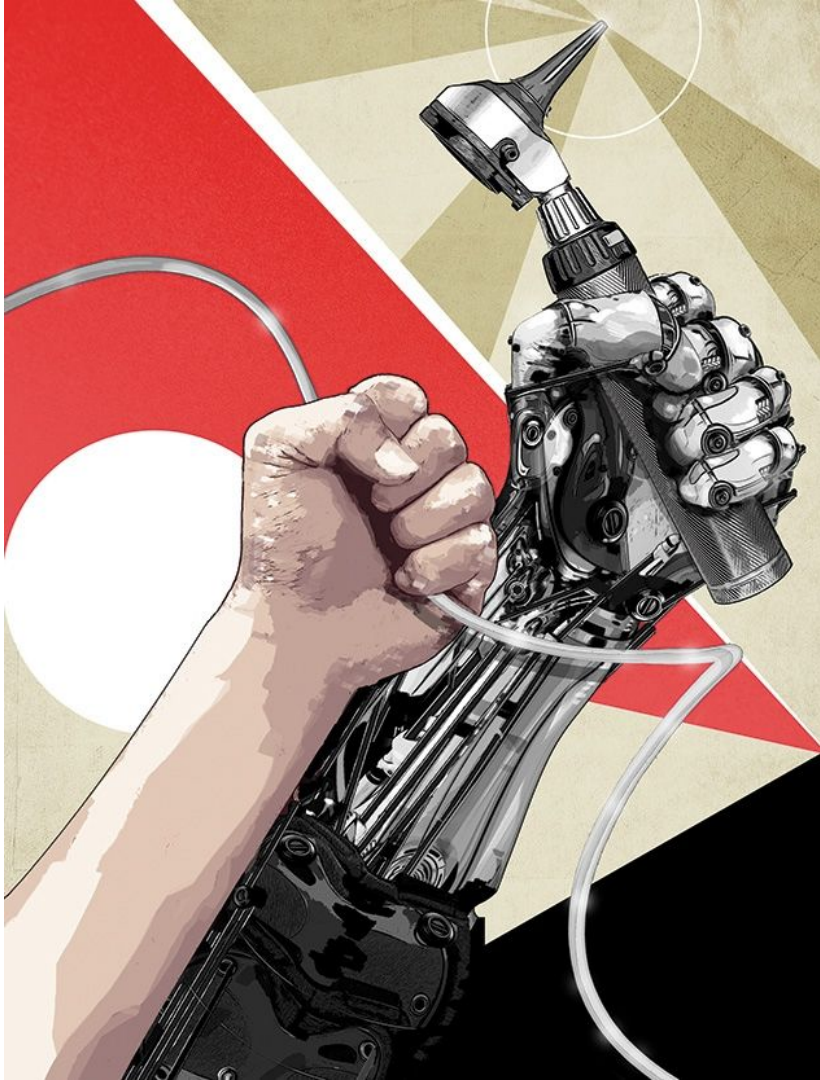
Next session:
REGRESSION

Saturdays.AI

Practice - EDA!



Challenge!



Bibliografía

Saturdays.AI

/1./ /Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow/

/2./ /Fast.AI - Introduction to Machine Learning for Coders/

/3./ /MLCourse.AI/

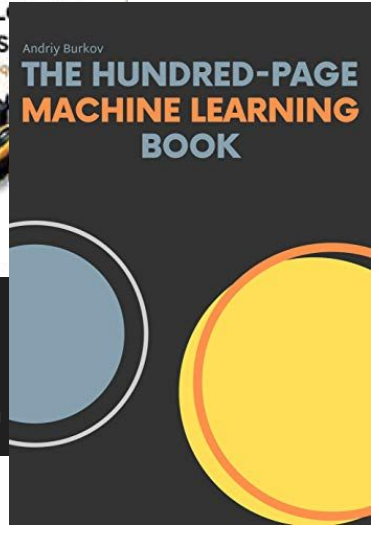
/4./ /DeltaAnalytics/

/5./ /The Hundred-page Machine Learning Book/

/6./ /Machine Learning for Humans (Vishal Maini)/

/7./ /Datacamp/

/8./ /DataQuest/



Partners

Agradecemos a nuestros partners por confiar en nosotros para facilitar la formación en IA de cara a la 4ª Revolución Industrial.